DOCUMENT RESUME

ED 208 025 TM 810 717

AUTHOR Fitzpatrick, Anne R.
TITLE Validating Decisions Made with Criterion-Referenced Tests.
PUB DATE Apr 81
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Classification; Competence; *Criterion Referenced Tests; Cutting Scores; Evaluation Methods; Predictor Variables; *Test Results; *Test Use; *Test Validity

ABSTRACT

Three kinds of classificatory decisions that might be made using criterion-referenced tests (CRTs) are described, and methods to appraise the validity of each are subsequently discussed. Decisions that entail predictive, descriptive, and evaluative classifications comprised the three kinds of decisions described. Predictive classifications entail assigning individuals who take a test to categories that are intended to predict their status on a criterion variable. It is suggested that these decisions can be appraised for validity using forms of the familiar criterion-related validity study. Considered as examples of descriptive decisions are the mastery and competency decisions that are commonly made using CRTs. Evaluative classification decisions would be made when judges set a passing score regarded as desirable to require. When these performance classifications reflect judges' intrinsic valuations of different levels of test performance, it is argued that the validity of these classfications cannot be shown. However, when these performance classifications reflect judges' instrumental valuations of different levels of test performance, their validity should be shown by criterion-related validity studies. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED208025

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Validating Decisions Made with Criterion-Referenced Tests[1]

Anne R. Fitzpatrick[2]
University of Massachusetts, Amherst

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
A.R. Fitzpatrick
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 717

[1]A paper presented at the annual meeting of the American Educational Research Association, Los Angeles, 1981.

[2]The author thanks Ronald K. Hambleton, Elaine Kachavos, and Robert Simon for helpful comments on a draft of this paper.

Abstract

In this paper, three kinds of classificatory decisions that might be made using criterion-referenced tests (CRTs) were first described, and then methods that could be used to appraise the validity of each of these kinds of decisions were subsequently discussed. Decisions that entail predictive, descriptive, and evaluative classifications comprised the three kinds of decisions that were described. Predictive classifications were said to entail assigning individuals who take a test to categories that are intended to predict their status on a criterion variable. It was suggested that these decisions can be appraised for validity using forms of the familiar criterion-related validity study. Considered as examples of descriptive decisions were the mastery and competency decisions that are commonly made using CRTs. When it is presumed that mastery or competency classifications characterize the status of classified individuals' knowledge, skill, ability, or the like, it was argued that these classifications are properly regarded as referring to levels of a construct and as warranting construct validation to establish the degree to which they are accurate. With regard to evaluative classifications, it was said that this kind of decision would be made when judges set a passing score that they regard as desirable to require; then levels of performance above and below this score represent classes of test performance that judges regard as acceptable and unacceptable, respectively. When these performance classifications reflect judges' intrinsic valuations of different levels of test performance, it was argued that the validity of these classifications can not be shown. However, when these performance classifications reflect judges' instrumental valuations of different levels of test performance, it was said that their validity should be shown by criterion-related validity studies.

Validating Decisions Made with Criterion-Referenced Tests

Test validity has been said to refer to the degree to which test score interpretations or uses are shown to be accurate (Cronbach, 1971; Linn, 1979). When criterion-referenced tests (CRTs) are employed, one of two kinds of test score interpretations or uses is likely to be associated with the scores that individuals obtain on these tests (Hambleton, Swaminathan, Algina & Coulson, 1978). On the one hand, examinees' scores on a CRT may be regarded as descriptors of what they have learned or of what they are skilled to do. On the other hand, CRT scores may be used as a basis for classifying examinees. In the first of these two situations, validity concerns the degree to which it is accurate to infer individuals' level of knowledge or skill from their levels of test performance. In the second situation, validity concerns the degree to which it is accurate to infer individuals' classifications from their levels of test performance. In any situation, this validity would be established by carrying out empirical investigations to gather evidence that shows the accuracy of the meaning or use one wishes to ascribe to the set of test scores being considered (Campbell & Fiske, 1959; Messick, 1975).

The question of what methods should be used to gather evidence that establishes the validity of descriptive uses of CRT scores has

been the subject of considerable deliberation among test specialists (e.g., Linn, 1979; Messick, 1975; Millman, 1977), but to date there has not been much discussion of what procedures should be used to examine the validity of using CRT scores to make the various kinds of classificatory decisions that interest CRT practitioners. Although Linn (1979) and others have noted that it is important to establish the validity of these decisions to ensure that examinees are classified properly, specialists who have been concerned with the use of CRTs for decision-making have primarily focussed their discussions on issues related to assessing the reliability rather than the validity of these decisions (e.g., Berk, 1980; Huynh & Saunders, 1980).

The offering in this paper is a discussion of methods that can be used to examine the validity of various classificatory decisions that are of interest in criterion-referenced testing contexts. In the pages that follow, three types of classificatory decisions that are commonly made in CRT contexts are first described. Subsequently noted are what we regard to be appropriate procedures that CRT practitioners can use to investigate the accuracy of each of these kinds of decisions. It should be noted that selection decisions are not treated in the following pages because there is already much discussion about methods for validating these decisions that is being conducted elsewhere (e.g., Dunnette & Borman, 1979; Ebel, 1977; Guion, 1976; Tenopyr, 1977).

Types of Decisions Made with Criterion-Referenced Tests

One type of decision that is made using CRTs entails assigning examinees on the basis of their test performance to classifications that are intended to be predictive of their performance on a criterion measure. For example, when units of a curriculum are arranged in a hierarchy, a cut-score may be set on a unit test in the interest of distinguishing between students who will and will not successfully grasp the material of a subsequent unit of the curriculum (Huynh & Perney, 1979; Livingston, 1976). In this circumstance, it is inferred that performance at or above this cut-score is predictive of successful performance on the subsequent unit and that performance below this cut-score is predictive of non-success on the subsequent unit of the curriculum. In another circumstance, say, where a basic skills math test is used, it may be inferred that examinees who perform above and below a cut-score that has been set on this test will perform well and poorly, respectively, in math courses they are offered in subsequent grades.

When decisions made using a test entail assigning examinees to classifications that are intended to be predictive of their status on a criterion variable, it seems appropriate to refer to these decisions as "predictive classifications" since predictive inferences are associated with these classifications. When criterion-referenced tests are used, these classifications typically are dichotomous and are intended to predict examinees' status on a dichotomous criterion variable (Hambleton, 1980).

A second kind of decision that is made using CRTs seems to entail classifications that are accompanied by descriptive inferences. Most notably, it would seem that descriptive inferences may be drawn when individuals are designated on the basis of their test performance as "masters" and "non-masters" or as "competent" and "not competent" in the subject matter or skill area that is covered by the test. Although some specialists have operationally defined these designations (e.g., Bloom, 1976; Bunda & Sanders, 1979), other specialists have suggested that these terms characterize the state of classified individuals' knowledge, skills, and/or capabilities. For example, Davis and Diamond (1974) have said that mastery means "complete knowledge, skill, or control" (p. 133). Other researchers have suggested that mastery refers to achievement at the upper end of a knowledge, skill, or ability continuum (Glaser, 1963; Meskauskas, 1976). With regard to the notion of competency, Finch (1980) stated that "the ability to use a specific skill or set of skills to meet the needs of a specific situation" (p. 400) constitutes competency in an area. Klemp (1979) defined competency to be "a generic knowledge, skill, self-schema or motive of a person" (p. 42) that produces what is publicly regarded as effective performance of diverse tasks. Presumably, non-competency as well as non-mastery would be defined by these specialists as a lack of those qualities that are thought to constitute, respectively, competency and mastery (e.g., Macready & Dayton, 1977).

Although it is the case that the meaning of mastery has not been agreed upon (Skager, 1974), and that the meaning of competency

is only currently being considered (e.g., Bunda & Sanders, 1979; Miller, 1978; Pottinger & Goldsmith, 1979), the above-mentioned views suggest that inferences about the state of individuals' attributes and/or capabilities may be drawn when these individuals are classified as "competent" and "not competent" or as "masters" and "non-masters" in light of their performance on a test. When mastery, competency, or other classifications are accompanied by such inferences, it would seem that they can aptly be called "descriptive classifications."

A third kind of decision that is made using CRTs appears to entail assigning examinees to what we would call "evaluative classifications." To grasp the nature of this kind of decision, consider the meaning of classifications that will be made using a passing score that has been set on, say, a basic skills test and reflects the level of test performance that judges think a high school senior ought to show on this test (Science Research Associates, 1966). Similarly, consider the classifications that will be made using a cut-score that reflects the passing score that judges think will effect certain pass/fail rates and/or educational, social, or political consequences that they regard as beneficial (e.g., Hambleton, 1980; Jaeger, 1978; Millman, 1973; Shepard, 1976). The cut-score that results when either of these judgmental procedures are used can be said to represent the passing score that judges regard as desirable to require. Also, the levels of test performance at or above this cut-score can be said to constitute a class of what judges regard as "acceptable" test performance, and the levels of test performance below this score can be

said to constitute a class of what judges regard as "unacceptable" test performance. Examinees taking a test on which such a cut-score has been set will therefore be assigned to classifications that describe the acceptability of their test performance. Because acceptability is a matter that is determined by individuals' values (Kaplan, 1964; Weber, 1949), it seems appropriate to characterize such classifications as "evaluative" in nature.

Methods for Validating Types of CRT Decisions

The Validation of Predictive Classifications

With respect to the decisions that are made with CRTs and entail predictive classifications, test specialists have indicated that forms of the familiar criterion-related validity study can be used to appraise the validity of these decisions. Specifically, studies of the relations between test and criterion classifications have been recommended as means of ascertaining the accuracy of inferring that individuals' test classifications predict their level of criterion performance (Hambleton, 1980; Huynh & Perney, 1979; Livingston, 1976). To conduct such studies, typically it is suggested either that measures of individuals' test and criterion performance be obtained, or that the known groups method be used, in which case the test under study would be administered to groups of individuals who are known to represent those levels of the criterion variable that the classifications made using the test are intended to predict. To the extent that one finds from either of these studies that individuals' test

and criterion classifications correspond in the manner intended, evidence is gained that these classifications can be validly regarded as predictors of individuals' levels of criterion performance (Berk, 1976; Hambleton, 1980).

As an illustration of such studies, consider the circumstance in which pass/fail classifications that are made with a criterion-referenced, first-year algebra test are intended to be predictive of, respectively, success and non-success in a second-year algebra course. The validity of making these classifications on the basis of the pass/fail cut-score that has been set on this test could be ascertained by appraising the relations between examinees' test classifications and their level of performance in the second-year algebra course. Alternatively, the validity of these classifications could be appraised by examining the test performance of students who are currently enrolled in the second-year algebra course and are known to be course successes and non-successes. In either case, the accuracy of one's pass/fail classifications would be indicated by the degree to which it is found that examinees who exceed the cut-score set on the algebra test are successes in the second-year algebra course, and that examinees who fall short of this cut-score do not perform successfully in the second-year algebra course.

The Validation of Descriptive Classifications

Suggested by the views that were cited above is the notion that both mastery and competency classifications should be thought of as referring to levels of a construct and as warranting construct

validation when these classifications are thought to describe classified individuals in some way. As will be recalled, it was noted that specialists using these classifications to describe individuals have indicated that they ascribe particular levels or kinds of knowledge, skill, ability or the like to individuals who are deemed to be "masters" or "competent" in a certain subject matter or skill area. It was also suggested that this implied that these specialists would ascribe a lack of knowledge, skill, ability, or the like to individuals who are deemed to be "non-masters" or "not competent" in the subject matter or skill area of interest. Knowledge, skill, and ability have been noted in measurement literature to be unobservable attributes that are properly regarded as constructs (Cronbach, 1971; Messick, 1975). When the designations of "master" and "non-master" or "competent" and "not competent" are thought to refer to the status of such attributes, these designations then are properly regarded as terms that refer to the status of constructs. It must also then be said that a claim that individuals' test classifications reflect their levels of mastery or competency represents an inference about the state of unobservable attributes that these individuals possess, and that construct validation is warranted to establish the degree to which this inference is accurate. As is well-known, it is the process of construct validation that enables one to gain evidence of the accuracy of inferring that individuals' scores on a test reflect levels of an unobservable attribute, trait, or internal process (Messick, 1975; Millman, 1978).

The recommendation that construct validation should be used to establish the accuracy of mastery and competency decisions is at variance with the most commonly espoused views on proper validation of these decisions. Typically, test specialists have suggested that criterion-related validity studies of the kind mentioned previously should be conducted to appraise their validity (Berk, 1976; Huynh & Perney, 1979; Millman, 1974; Livingston, 1976; Zieky & Livingston, 1977). For example, Livingston (1976) and Huynh and Perney (1979) have indicated that the validation of mastery classifications could proceed by appraising classified individuals' performance on a criterion measure. Thus, individuals who have been classified as masters and non-masters on, say, a pencil-and-paper test could be asked to take a performance test on which success is presumed to reflect mastery and a lack of success is presumed to reflect non-mastery. According to these researchers, the mastery/non-mastery classifications made using the paper-and-pencil test could be considered valid to the extent that the examinees who exceeded the cut-score set on this test are found to be successful on the performance test, whereas the examinees who fell short of this cut-score are found to be unsuccessful on the performance test. In lieu of a criterion measure, Millman (1974) and Zieky and Livingston (1977), among others, discussed using criterion groups in an appraisal of the accuracy of mastery or competency classifications. To assess the accuracy of mastery classifications, these researchers suggested that an investigator could administer the mastery test of interest to two groups of individuals, only one of

which has been instructed in the subject matter that is covered by the test. They indicated that the decisions made with the mastery test could be considered valid to the extent that the individuals who have been instructed in the subject matter covered by the mastery test and found to exceed the cut-score that has been set on this test, whereas the individuals who have not been instructed in this subject matter are found to fall short of the cut-score when they take the test.

Our view that construct validation rather than a single criterion-related study is needed to establish the accuracy of mastery or competency decisions stems from the supposition that mastery and non-mastery as well as competency and non-competency designations typically will be associated with meanings that cannot be appraised for accuracy using a single criterion-related validity study. Since constructs have no definite criterion measure unless they are operationally defined (Cronbach & Meehl, 1955), simply to regard mastery and competency designations as terms that refer to levels of unobservable knowledge, skill, and/or ability constructs seemingly would preclude identifying a single criterion that can be regarded as a fully valid indicator of individuals' mastery or competency levels. Moreover, it is likely that both mastery and competency levels will be ascribed with other meanings that cannot be confirmed by a single criterion-related validity study. For example, both mastery and competency have been conceived of as attributes that can be affected by instruction, since it is commonly suggested that effective instruction will improve the mastery or competency status of individuals who have been deemed to be non-masters

or not competent in a particular subject matter or skill area (Linn, 1979; Sallander, 1980). Levels of mastery have also been linked with individuals' ability to retain and transfer what they have learned (Block, 1972), and levels of competency have been said to affect individuals' subsequent academic success (Torshen, 1977), job success (Klemp, 1979), or their success in solving the problems that are posed in everyday life (Haney & Madaus, 1978). When mastery and non-mastery or competency and non-competency labels are ascribed with such meanings, the diverse investigations that are entailed in construct validation would be needed to explore the several relations between these classifications and the factors or variables with which levels of competency or mastery have been linked.

The particular procedures that one would use to investigate the construct validity of descriptive mastery or competency classifications would be selected in light of one's ideas or hypotheses about the meaning of mastery and non-mastery or the meaning of competency and non-competency in the subject matter or skill area that is to be tested. These ideas or hypotheses would be comprised of predictions like those noted above, which specify how levels of mastery or levels of competency should relate to and affect, or be affected by, certain variables, factors, and/or other constructs (Cronbach, 1971; Cronbach & Meehl, 1955). For example, one might posit that individuals' competency and non-competency in, say, algebra skills should not be strongly related to their math aptitude, but should be affected by whether they have or have not grasped certain concepts and rules of arithmetic and geometry. It might also be proposed that competency

in algebra can be gained from effective instruction and that this competency should enable individuals to successfully solve everyday problems that entail algebraic calculations.

Once one's ideas about the meaning of mastery or competency have been formulated, data can then be gathered that will enable one to establish whether the classifications to be made with one's test have this meaning. Observations should be taken that will show what relations exist between the classifications that are made using one's mastery or competency test and those factors, variables, and/or constructs that one has linked with levels of mastery or competency. Also, studies should be conducted to gain data that bear on plausible counterinterpretations that might be offered to explain the meanings of these classifications (e.g., Campbell, 1960; Cronbach, 1971). To the extent that one finds from such studies that one's test classifications show the pattern of relations that only one's notion about mastery or competency would explain, support is secured for the validity of one's claim that these classifications reflect individuals' level of mastery or competency, whichever one intends (Messick, 1975).

As an illustration of the kind of construct validation study that is being recommended, let us say that we are to formulate a construct validity study in the interest of examining the accuracy of a claim that junior high school students' competency and non-competency in algebra skills is indicated by the classifications that they have been assigned in light of their performance on a test that has been devised. Let us also assume that the above-mentioned perspective constitutes the theory that we hold about the meaning of

competency in these skills. Since this perspective indicates that we should expect certain relations to be obtained between individuals' levels of competency in these skills and both their math aptitude and their grasp of certain concepts and rules of arithmetic and geometry, we might first examine the correlations between junior high school students' test classifications and a measure of their math aptitude, and then examine the correlation between these classifications and the students' grades in, say, their fifth and sixth grade math classes where the mathematical concepts and rules of interest have been taught. Since we also expect that individuals' level of competency in algebra should affect how well they solve everyday problems that require these skills, we could also examine the correlation between these students' test classifications and ratings of their success or non-success in performing tasks that pose these kinds of problems. Our supposition that instruction will effect competency should also be examined: We could ask students who have been designated as not competent in light of their performance on our competency test to retake this test after they have been tutored in the algebra skills that they have been presumed to lack. Finally, to rule out the counterinterpretation that our test classifications reflect individuals' mastery rather than their competency in algebra skills, we could calculate the correlation between students' test classifications and teachers' ratings of their mastery status with regard to these skills.

The findings from construct validity studies like the one that has been illustrated here provide one with bases for not only

confirming but also for improving the accuracy of one's mastery or competency test classifications. To the extent that these findings confirm one's predictions about the relations between these classifications and the other measures that one has taken, evidence is gained that these classifications have the meaning one intends. The findings that one obtains might also indicate that the accuracy of one's predictions could be improved by adjusting the cut-score that has been set on one's mastery or competency test, so that an adjustment of this cut-score would improve the accuracy of the test classifications that are made. Finally, these findings might indicate that one's mastery or comeptency test classifications bear much weaker relations than expected with other measures that one has taken, in which case one should not only question the validity of inferring individuals' mastery or competency status from these classifications, but also one should reconsider and perhaps modify one's theory about the meaning of mastery or competency levels that has been used to guide one's investigations (Cronbach, 1971; Cronbach & Meehl, 1955).

The Validation of Evaluative Classifications

When a cut-score that is set on a test reflects the passing score that judges regard as desirable to require, it was suggested earlier that this cut-score distinguishes between classes of what judges deem to be acceptable and unacceptable test performance. Because acceptability is a matter determined by individuals' values, it was suggested that such classifications should be characterized as evaluative in nature.

To determine what methods can be used to appraise the validity of classifying examinees in terms of the acceptability of their test performance, it is helpful to first consider what is meant by these classifications. In some circumstances, judges' designations of what are acceptable and unacceptable levels of performance might represent simply what they consider to be good and poor, or desirable and undesirable, levels of test performance. In this case, it can be said that these designations reflect judgments about the intrinsic value of different levels of test performance: The levels of this performance deemed acceptable are test results that have been judged to be good or desirable in themselves, and the levels of this performance deemed unacceptable are test results that have been judged to be poor or undesirable in themselves (e.g., Moore, 1922). Such meanings can reasonably be ascribed to the classifications that will be made using a cut-score that has been set, say, on a basic skills test and is intended to reflect the level of performance that judges think a high school senior ought to attain or exceed (Science Research Associates, 1966). Here it can be said that performances above and below such a cut-score are what judges regard as respectively, desirable and undesirable test results for a high school senior to obtain.

In other circumstances, judges' valuations of test performance as acceptable and unacceptable might reflect their judgments not of the intrinsic value of this performance, but rather of the instrumental value of this performance. That is, they deem some levels of test performance as acceptable because they associate these levels with effects or consequences that they regard as good or desirable, and they

deem other levels of performance as unacceptable because they associate these levels with effects or consequences that they regard as not good or not desirable (e.g., Moore, 1922). These meanings could be ascribed to test classifications that are made using a cut-score intended to effect certain pass/fail rates, or certain educational, political, or social outcomes (Jaeger, 1976; Millman, 1973; Shepard, 1976). For instance, the classes of performance distinguished by such a cut-score might be deemed acceptable and unacceptable because these performances are presumed to lead to later academic success and non-success, respectively (Millman, 1973).

Given that the classifications of test performance as acceptable and unacceptable reflect value judgments about this performance that are either intrinsic or instrumental in nature, what can be said about the manner in which these classifications can be validated? When these classifications reflect judgments about the intrinsic goodness or desirability of test performance, we would contend that they cannot be appraised for or said to have any kind of test validity. This is because there is no means of empirically demonstrating that such value judgments are valid. The validity of a score interpretation or use rests on the degree to which evidence that is gained from independent observations shows that the set of scores under study has the meaning or use that one intends (Campbell & Fiske, 1959; Cronbach, 1971). For example, the validity of a claim that a set of scores reflects individuals' reading comprehension in part can be ascertained from the correlation between these scores and another measure of their reading comprehension that is taken (Campbell & Fiske, 1959). In contrast, when it is claimed that individuals' performance on, say,

a reading test is acceptable or unacceptable, there are no observations related to this performance that can be taken and used as a basis for inferring the acceptability of this performance. Hence, the validity of this claim cannot be shown. To grasp this point fully, note that we might take measures of several properties of a color, but none of these measures would indicate that this is a color that we particularly like or dislike. In short, it is not possible to acquire data from which we can infer the validity of judgments about the intrinsic acceptability or desirability of different levels of test performance.

When test performance is classified as acceptable and unacceptable, and these classifications reflect judgments about the instrumental value of this performance, we would contend that the validity of these classifications can and should be shown. To see the basis for this contention, consider as an example the circumstance in which certain levels of performance on a basic math skills test are deemed acceptable because it is thought that these skill levels are needed to successfully solve certain important problems that arise in everyday life. Concomitantly, other levels of performance on this test are deemed to be unacceptable because it is thought that these skill levels will not enable success in solving such everyday problems. In this situation, predictions form the basis for these valuations of test performance. That is, the levels of performance judged to be acceptable are those that have been predicted to effect success in solving everyday math problems, and the levels of performance judged to be unacceptable are those that have been predicted to effect non-success in solving these problems. Since observations can be taken

that will show the accuracy of these predictions, it is possible to demonstrate the validity of such valuations. Specifically, this validity can be shown by criterion-related studies designed to appraise the relations between individuals' test classifications and their scores on measures of the outcomes that this performance is expected to affect. Accordingly, to appraise the validity of classifying examinees' performances on the above-mentioned math skills test as acceptable and unacceptable, one could appraise the relation between examinees' test classifications and their performance on a certain set of everyday problems since it is anticipated that those examinees said to show acceptable test performance will successfully solve these problems, whereas those examinees said to show unacceptable performance will not experience this success. Validity in classifying examinees' test performance as acceptable and unacceptable can be claimed to the extent that one finds from such a study that there is the predicted relation between these classifications and the good and poor outcomes with which these classifications have been associated.

Summary

The purpose of this paper was to identify the validation procedures that CRT practitioners can use to show the validity of classificatory decisions that are commonly of interest in circumstances where criterion-referenced tests are used. In the paper, three types of classificatory decisions that might be made using CRTS were first

briefly described. Subsequently, methods were noted that could be used to appraise the validity of each of these kinds of decisions.

Decisions that entail predictive classifications comprised the first type of decision that was described. Since this kind of decision entails assigning individuals who take a test to categories that are intended to predict their status on a criterion variable, it was indicated that these decisions could be appraised for validity using forms of the familiar criterion-related validity study.

Decisions that entail what was called descriptive classifications comprised the second type of decision that was described in the text above. Considered as examples of this type of decision were the mastery and competency classifications that are commonly made with CRTs. It was argued that when it is presumed that the master/non-master or competent/not competent classifications characterize the state of classified individuals' knowledge, skill, ability, or the like, these classifications are properly regarded as referring to levels of a construct and as warranting construct validation to establish the degree to which they are accurate. The manner in which a construct validity study of mastery or competency classifications should be formulated was briefly discussed and an example of such a study was provided.

Decisions that entail what were regarded as evaluative classifications comprised the third type of decision that was treated in the text above. This kind of decision is made using a cut-score that represents the passing score that judges think is desirable to require. It was said that levels of performance above and below this

score can be regarded as classes of acceptable and unacceptable performance, respectively, so that examinees who take a test on which such a cut-score has been set will be classified in terms of the acceptability of their test performance. It was argued that the validity of classifying examinees' performance as acceptable or unacceptable cannot be shown when these classifications reflect judgments of the intrinsic value of this performance. However, when these classifications reflect judgments of the instrumental value of this performance, it was said that their validity can and should be shown.

References

Berk, R. A. Determination of optimal cutting scores in criterion-referenced measurement. Journal of Experimental Education, 1976, 45, 4-9.

Berk, R. A. A consumer's guide to criterion-referenced test reliability. Journal of Educational Measurement, 1980, 17, 323-350.

Block, J. H. Student evaluation: Toward the setting of mastery performance standards. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

Bloom, B. S. Human characteristics and school learning. New York: McGraw-Hill, 1976.

Bunda, M. A., & Sanders, J. R. (Eds.) Practices and problems in competency-based measurement. Washington, DC: National Council on Measurement in Education, 1979.

Campbell, D. T. Recommendations for APA test standards regarding construct, trait or discriminant validity. American Psychologist, 1960, 15, 546-553.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Caws, P. The philosophy of science. Princeton, NJ: D. Van Nostrand, 1965.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington, DC: American Council on Education, 1971.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

Davis, F. B., & Diamond, J. J. The preparation of criterion-referenced tests. In C. H. Harris, M. C. Alkin, & W. J. Popham (Eds.), CES Monograph Series in Evaluation: Problems in criterion-referenced measurement. Los Angeles, CA: University of California Center for the Study of Evaluation, 1974.

Dunnette, M. D., & Borman, W. C. Personnel selection and classification systems. Annual Review of Psychology, 1979, 30, 477-525.

Ebel, R. L. Comments on some problems of employment testing. Personnel Psychology, 1977, 30, 55-63.

Finch, F. A taxonomy for competency testing programs. In R. M. Jaeger, & C. K. Tittle (Eds.), Minimum competency achievement testing. Berkeley, CA: McCutchan, 1980.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Guion, R. M. Recruiting, selecting and job placement. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally, 1976.

Hambleton, R. K. Test score validity and standard-setting methods. In R. Berk (Ed.), Criterion-referenced measurement: State of the art. Baltimore, MD: Johns Hopkins Press, 1980.

Hambleton, R. K. Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Haney, W., & Madaus, G. Making sense of the competency testing movement. Harvard Educational Review, 1978, 48, 462-484.

Huynh, H., & Perney, J. C. Determination of mastery scores when instructional units are linearly related. Educational and Psychological Measurement, 1979, 39, 317-325.

Huynh H., & Saunders, J. C. Solutions for some technical problems in domain-referenced mastery testing: Final report. Columbia, SC: Department of Educational Research and Psychology, College of Education, University of South Carolina, August, 1980.

Jaeger, R. M. A proposal for setting a standard on the North Carolina High School Competency Test. Paper presented at the 1978 Spring meeting of the North Carolina Association for Research in Education, Chapel Hill, 1978.

Kaplan, A. The conduct of inquiry. San Francisco, CA: Chandler, 1964.

Klemp, G. O., Jr. Identifying, measuring, and integrating competence. In P. S. Pottinger, & J. Goldsmith (Eds.), New directions for educational measurement (Vol. 3): Defining and measuring competence. San Francisco, CA: Jossey-Bass, 1979.

Linn, R. L. Issues of validity in measurement for competency-based programs. In M. A. Bunda, & J. R. Sanders (Eds.), Practices and problems in competency-based measurement. Washington, DC: National Council on Measurement in Education, 1974.

Livingston, S. A. Choosing minimum passing scores by stochastic approximation techniques. Report No. COPA-76-02. Princeton, NJ: Center for Occupational and Professional Assessment, Educational Testing Service, 1976.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. Review of Educational Research, 1976, 46, 133-158.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

Miller, B. S. Minimum competency testing: A report of four regional conferences. St. Louis, M: CEMREL, Inc., 1978.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, CA: McCutchan, 1974.

Millman, J. Strategies for constructing criterion-referenced assessment instruments. Paper presented at the Conference on Large-Scale Assessment, Denver, 1978.

Moore, G. E. Principia ethica. London: Cambridge University Press, 1922.

Pipho, C. Minimum competency testing in 1978: A look at state standards. Phi Delta Kappan, 1978, 59, 585-588.

Pottinger, P. S., & Goldsmith, J. S. Editor's comments. In P. S. Pottinger, & J. S. Goldsmith (Eds.), New directions for experiential learning (Vol. 3): Defining and measuring competence. San Francisco, CA: Jossey-Bass, 1979.

Sallander, R. Competency tests: Decisions for educators. In R. M. Jaeger, & C. K. Tittle (Eds.), Minimum competency achievement testing. Berkeley, CA: McCutchan, 1980.

Science Research Associates. ITED. The Iowa Tests of Educational Development: Manual for the school administrator. Chicago: Science Research Associates, 1966.

Shepard, L. Setting standards and living with them. Florida Journal of Educational Research, 1976, 18, 23-32.

Skager, R. W. Generating criterion-referenced tests from objectives-based assessment systems: Unsolved problems in test development, assembly, and interpretation. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. Los Angeles, CA: University of California Center for the Study of Evaluation, 1974.

Tenopyr, M. L. Content-construct confusion. Personnel Psychology, 1977, 30, 47-54.

Torshen, K. P. The mastery approach to competency-based education. New York: Academic Press, 1977.

Weber, M. On the methodology of the social sciences. Translated and edited by E. A. Shils & H. A. Finch. New York: Free Press, 1949.

Zieky, M. S., & Livingston, S. A. Manual for setting standards on the Basic Skills Assessment Tests. Princeton, NJ: Educational Testing Service, 1977.